MANAV RACHNA INTERNATIONAL UNIVERSITY

Research Plan
For

Ph. D. Programme 2009-10

Title: Design and Development of New Algorithm for Web Content Mining Suitable for Mobile Systems

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:
Name: Neha Gupta
Registration No: 09019990231

Supervisor:

Name: Dr. Babu Ram
Designation: Professor
Department of Computer Applications
Manav Rachna International University
Faridabad

Co-Supervisor:

Name: Dr. Saba Hilal
Designation: Director
Fausta Research & Development Ltd,
Faridabad

ABSTRACT

The Problem of Kids based secure internet search is grabbing attention from every facet of society. A detailed study of existing approaches and software will be provided. The denigration is on the limited depth of existing models and techniques to provide Secure and Relevant access to kids according to their age. During the course of the study we will develop new algorithm and technique to extract the secure and fast educational content for kids according to their age and redirect the interest of the kids towards the educational aspect of the query without blocking or hiding anything from them. The research work will focus on 4 areas while mining the web for secure and quality data: first is the use of web mining in extracting educational content, second is child query difficulties and their needs, third is child security, and fourth is content adaptation on mobile devices. The initial proposals require the detailed study and analysis of Kids query patterns, their difficulties, the security concern and the need of the new interface.

The research work will demonstrate the use of various web mining principals, Educational quality specifications, Query enhancement techniques, kids search patterns and their difficulties and the various content adaptation techniques.

Keywords: Query Enhancement, Web content Mining, Kids, Search Pattern, Content Adaptation, Internet, Educational Content

Table of Contents

S.No	Description	Page No.
1.0	Introduction	4
2.0	Literature Review	5-6
3.0	Description of Broad Area 3.1 Web Content Mining for Kids	7-8
4.0	Objective of the study	8
5.0	Methodology to be adopted	9
6.0	Proposed Solution	10
7.0	References	11 - 12

1.0 Introduction

If there is one area of certain growth for tomorrow, it is the mobile communications sector. And with significant improvements of technology over the last few years in vital areas such as data transfer speed, screens and content, mobile devices are challenging the PC as an Internet access device of choice. This in turn has raised a key issue in the Mobile Internet space: that of the extraction of concise and accurate data for mobile devices while searching various educational queries on mobile interface. [1].Usage of mobile has various issues including smaller screens, typing limitations of phone keypads and the cost of spending lots of time scrolling through mobile search results. [2].Students now a days are using mobile phones to access the internet for their school assignments but problem arises when the search engine (i.e. Google) give various links and students need to scroll a lot and search a lot for landing on the desired output. Second problem arises because of non-compatible mobile web sites and the format of data that a search engine displays on mobile device. To solve this problem an interface need to be developed which fill the gap between the mobile device and the search engine capabilities. For this gap to be filled various technical issues and concepts of web content mining and mobile web need to be explored and implemented. For example implementation and working of WAP gateway, WAP Protocol, .Net/ Java platform, Web servers functioning, Content mining editors, Data management etc.

Today the students expect constant access to information—in the classroom and beyond. That's why more and more students, faculty and life long learners are using Mobile devices to access information from the web. As educators all over the world are discovering mobile learning works so the interface need to design for the students to provide help in their school assignments and other search related works. The mobile interface will help the student to search a particular query on web and give the result in concise and accurate and secure form on the mobile device and if require redirect the interest of the child towards educational content only.

2.0 Literature Review

Researchers have explored ways to build systems that automatically gather and manipulate Web based information on a user's behalf. But as the relevant content is embedded in HTML pages, extracting their content is difficult. A wrapper is a procedure for extracting a particular resource's content. Kushmerick & Weld et al. [7] used Wrapper induction for information extraction and introduced wrapper induction, a technique for automatically constructing wrappers. Another related study by Crescenzi & et al. [5], describes a project RoadRunner to investigate techniques for extracting data from HTML sites through the use of automatically generated wrappers.

Butler & Liu et al. [2] Presents a fully automated extraction system for the World Wide Web, called Omini. Omini parses web pages into tree structures and performs object extraction. In another research paper Chang & Liu [3] describes information extraction based on pattern discovery. They present a technique that extracts information blocks without training examples using a data structure called a PAT tree. PAT trees allow the system to efficiently recognize repeated patterns in a semi-structured Web page.

Roesnfeld et al. [10] describe a general procedure for structural extraction, which allows for automatic extraction of entities from the document based on their visual characteristics and relative position in the document layout. Often fonts, physical positioning and other graphical characteristics are used to provide additional context to the information. This information is lost with pure text analysis. They also describe a specific implementation of the procedure to PDF documents, called PES (PDF Extraction System).

Yu & Cai et al. [14] proposed another study based on web content structure considering the visual representation. The paper specifies how a user understands web layout structure based on the visual perception and the scheme is independent of underlying documentation representation.

The structural data extraction work of Arasu et al. [1], focuses on web sites, containing large sets of pages generated using a common template or layout. They have studied the problem of automatically extracting the database values from such template-generated web pages without any learning examples or other similar human input.

Gupta & Kaiser et al. [6] uses DOM based Content Extraction of HTML documents; to clean the web pages from unimportant images and extraneous links around the body of an article that distracts a user from actual content. They have implemented this approach in a publicly available Web proxy.

Pinto & McCallum et al. [9], focused on the ability to find tables and extract information from them. This paper presents the use of conditional random fields CRFs for table extraction and compares them with hidden Markov models (HMMs). Wang & Lochovosky [13] focuses on scheme to help users query, extract and integrate data from web pages generated dynamically from databases, i.e., from the Hidden Web. They describe a system called, DeLa, which reconstructs (part of) a "hidden" back-end web database. It does this by sending queries through HTML forms, automatically generating regular expression wrappers to extract data objects from the result pages and restoring the retrieved data into a table.

Chriisment & Dousset et al. [4] work on extracting, exploring and visualizing geo-referenced information, presented a platform where the documents are analyzed in order to extract predefined elements such as location referenced. Then a multi-dimensional representation is extracted from each document and summarized in the form of tables from which the information is extracted. They combine mining and visualizing displaying the information graphically.

Liu and Zhai [8] presented automatic extraction of structured data from Web pages. Given a page, this method first builds a tag tree based on visual information. It then performs a post-order traversal of the tree and matches sub trees in the process using a tree edit distance method and visual information. It is based on finding data records and extracting items form them. The method works on both flat and nested data records.

Zhai & Liu [13] studied the extraction of Web data using instance based learning. In contract with the other approaches to data extraction including wrapper induction and automatic methods, they proposed an instance based learning method, which performed extraction by comparing each new instance to be extracted with labeled instances (or pages). The method works without an initial set of labeled pages to learn extraction rules as in wrapper induction and the algorithm is able to start extraction from a single labeled instance.

3.0 Description of Broad Area

3.1 Web Content Mining for Kids

For Kids searching the relevant content and weeding out the distracted content is the most challenging task as there is no fix rule to classify which content is appropriate for a kid who is searching to complete his homework and which is not. The amount of information on the Web is growing rapidly as well as the new users are inexperienced in the art of Web search [12]. It is estimated to contain approximately one billion pages of publicly accessible information and it continued to grow at an exponential rate, tripling in size over the past two years. Now the amount of information available on the net is unimaginable

There is so much data on the web that websites devoted to searching the internet, like Google, have developed. Search engines take queries and return results. Textual queries are the way a user describes the information he or she needs. The results in this case are web pages that the search engine infers contain the information the user's query described. For instance, a user might type in the query "Manav Rachna College" hoping to find information about Manav Rachna College. In this case, Google or any other search engine would return a list of pages that it found included the words "Manav Rachna College." Still there is no guarantee that the list would contain useful results and the user may have to go through pages of results to find the desired information. So imagine searching for information in the world's largest library (i.e. the web) where the books and journals (stripped of their covers and title pages) are shelved in no particular order and without reference to a central catalogue. A Search Engine helps to overcome these problems and navigate through the Internet to get the information that one requires quickly and accurately.

But for all its power, it is just now being tapped to transform education. Still, Internet provides a great platform for e-learning [11]. It includes Educational software, Programming languages, Educational Content Websites, School Websites, Virtual courses, Learning Management Systems, A-synchronous Learning Networks and Collaborative Learning Environments [11].

It points to the strong need of the tailor made services and content designed around the needs of the individual and that, which is available at a time and place and in a form, which suits the learner's needs [11]. The searching of educational information, content and material requires the development of better web content finding tools and techniques. Most of the times the required content is available on the Web but finding it is difficult. The Search Engines help to extract the information bundles from the vast ocean of the Web. However, finding of the correct collection still remains unsolved [11]. Moreover, most of the time

the search engine is not designed for the purpose that matches the user searches perspective. And above all if the content is found, the possibilities of distracting from the main content to the useless content is more.

Thus, there is an increasing need for research aimed at understanding children's information needs and to provide IR systems that suit the characteristics of content for children. Web content mining can be used to design an IR in such a way that kids remain focused on what they are searching, instead of distracting and jumping over to sites which may cause mental stress and let them lost in this big world.

4.0 Objective of the Study

All the above Papers discussed the way to extract the content from web pages either using wrappers or tree based approach, but security issues and relevant data extraction have not been taken into account.

Our approach is to develop an interface for the Students, Teachers and Life Long learners for their queries (educational or other) using internet through mobile devices taking care of security issues and relevant data extraction from the available data on the net.

The Interface will open the new doors of searching the web with the help of Mobile devices and also protect them from unwanted security risks and vulnerable data.

Security, Privacy, Quality & reliability of the content will be the main concern while searching and displaying the data on the mobile screen. Most of these systems are based on predefined databases however the proposed system will search the content on web (Educational or other), extract the relevant and accurate information, and fit it on small screen of mobile device. The Proposed interface will not be specific to a single knowledge base. Only Specific queries (educational or other) will be accepted by the system so that student will not be distracted to the vulnerable environment of web. The proposed system can handle the content belonging to a range of subjects.

5.0 Methodology to be adopted

We propose to Design and Develop new Algorithms for Web Content Mining suitable for mobile systems using required protocols and languages for searching various Internet sites and displaying the results in concise and accurate form on mobile systems. *The main aim of the research will be to develop an algorithm that will mine the web to extract the secure educational content for the kids' queries and redirect the interest of the child towards educational content and provides fast access to data in a secure way.* The focus will be on developing the concise, accurate and secure display of data on mobile devices using required protocols, Search Engines and various web content mining techniques. An interface compatible to mobile systems with added security features to access only the data related to him on mobile, is to be developed. This will keep the focus of the user his requirements and protect from the vulnerable environment of web.

The idea for developing this interface for mobile devices has many benefits.

Firstly mobile devices are widely used these days and are always online. So to perform a small search, student need not have to connect the laptop or pc to Internet. This mobile interface will give student an environment in which student can search the data anytime, anywhere and that too in secure way. Moreover, the benefits resulting because of the personalization features will also be there in the proposed mobile interface.

The research work will comprise the following modules:

Module 1: - Development of Mobile user Interface

Module 2: - Development of security controller

Module 3: - Development of algorithm for Web content Extractor

Module 4: - Development of Web content Presenter

Module 5: - Result Analysis & Evaluation

6.0 Proposed solution

The proposed solution for the problem stated above can be to develop an interface for a mobile device, through which student can access various internet sites in restricted manner for educational purpose only and can get accurate and concise result on mobile device in less time. The interface will be the implementation of required algorithm.

Proposed solution can be figured as:-

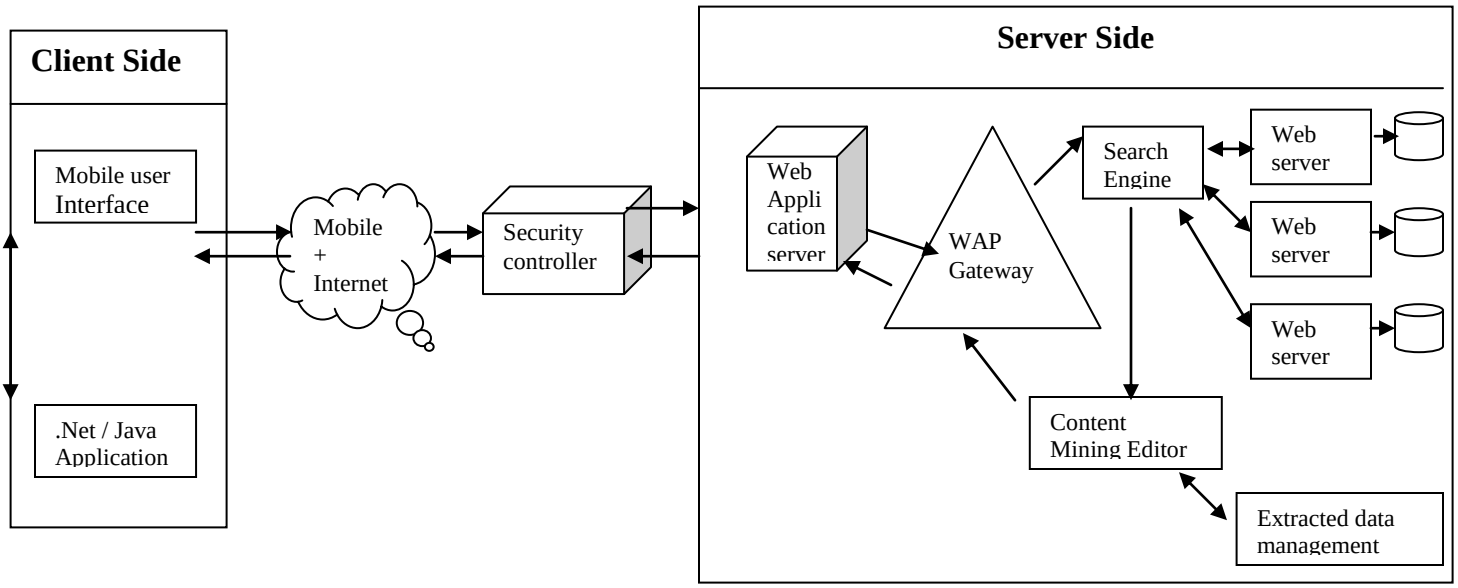

Figure 1:- Architecture of EWCM

Proposed solution covers the following points: -

- *A new Algorithm for extraction of secure educational content for Kids using Web content mining techniques.*
- *A new concept of redirection to be introduced.*
- *Redirecting the content extracted towards the educational content.*
- *Filtration of content on the basis of four content mining factors: Relevance, Quality, Reliability and Ease of Use.*

- *Development of MBEWCM to implement the algorithm and test the result.*
- *Adaptation of content for kids on mobile device.*
- *Helps in the improvement of quality and reliability of educational content.*
- *Testing the result with Google and Bing*

7.0 References

[1] Arasu, A., and Garcia-Molina, H., ***"Extracting Structured data from web Pages"***, ACM SIGMOD, pp 337-348 ,2003

[2] Buttler, D., Liu, L., Pu, C., "***A Fully automated extraction system for the the world wide web*** " , IEEE (9),pp 361-370, 2001

[3] Chang, C-H., Lui, S-L, ***"IEPAD: Information Extraction based on pattern discovery",*** ACM Digital Library , pp 681-688 , 2001

[4] Chriisment, C., Dousset, B,Karouach, S, Mothe, J., ***"Information mining: Extracting, Exploring and Visualizing geo-referenced information"*** , SIGIR-04 (Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval) ,Workshop on geographic information retrieval, 2004

[5] Crescenzi, V., Mecca, G., and Merialdo, P., "***Roadrunner : Towards automatic data extraction from large web sites***" , ACM Digital Library, pp 109-118, 2001

[6] Gupta , S., Kaiser, G., Neistadt, D. and Grimm, P., ***"DOM based content Extraction of HTML Documents",***ACM , pp 207-214,2003

[7] Kushmerick, N., Weld, D., and Doorenbos, R., "***Wrapper induction system for Information Extraction***" , Proc. of International Joint Conference on Artificial Intelligence ,1997

[8] Liu, B and Zhai, Y., NET- ***"A system for extracting web data from flat and nested data records"***, WISE-05 (Proceeding of 6th International Conference on Web Information system Engineering), 2005

[9] Pinto, D., McCallum, A., Wei, X. and Bruce, W., ***"Table extraction using conditional random Fields"***, ACM-SIGIR, pp 235-242 ,2003

[10] Rosenfeld, B., Feldman, R., Aumann, Y., ***"Structural extraction from visual layout of documents"***, Springerlink , pp 1-15, 2002

[11] Saba Hilal, S. A. M. Rizvi, "The Syllabus Based Web Content Extractor (SBWCE)", Communications of the IIMA, Vol 8 issue 1, 2008

[12] SK Dwivedi, "Word Sense Disambiguating Algorithm for Web Search Query Enhancement", International Conference on Cognitive Systems New Delhi, December 14-15, 2004

[13] Wang, J. and Lochovsky, F., ***"Data Extraction label Assignment for web databases"***, ACM, pp 187-196, 2003

[14] Yu,Cai,D,S., Wen, J-R and Ma, W-Y., ***"Extracting content structure for web pages based on visual representation"***, Springer – Verlag, 2642, pp- 406-417 , 2003

[15] Zhai, Y., and Liu, B., ***"Extracting web data using Instance- Based Learning"***, WISE-05 (Proceeding of 6th International Conference on Web Information system Engineering),2005